KARPOV.COURSES >>>
КОНСПЕКТ

> Конспект > 1 урок > Введение в Uplift-моделирование

> Оглавление

> Метрики
> Численные метрики

Ноутбук с практикой: скачать.

Данные: скачать.

Импортируемый модуль: скачать.

> Мотивация и постановка задачи

Чего хотим достичь?

Увеличения количества совершений целевых действий в вашем продукте.

> Типы воздействий на пользователя

Можно напоминать пользователю о себе:

Информация. Пример: “Сотни фильмов ждут тебя на Кинопоиске”.

Скидка или начисление баллов. Примеры:

- "Получите 200 баллов при покупке от 400 руб";
- "Вам начислено 300 баллов. Успейте списать их в течение следующей недели";
- "В 3 раза больше баллов в категории “Сосиски, сардельки".

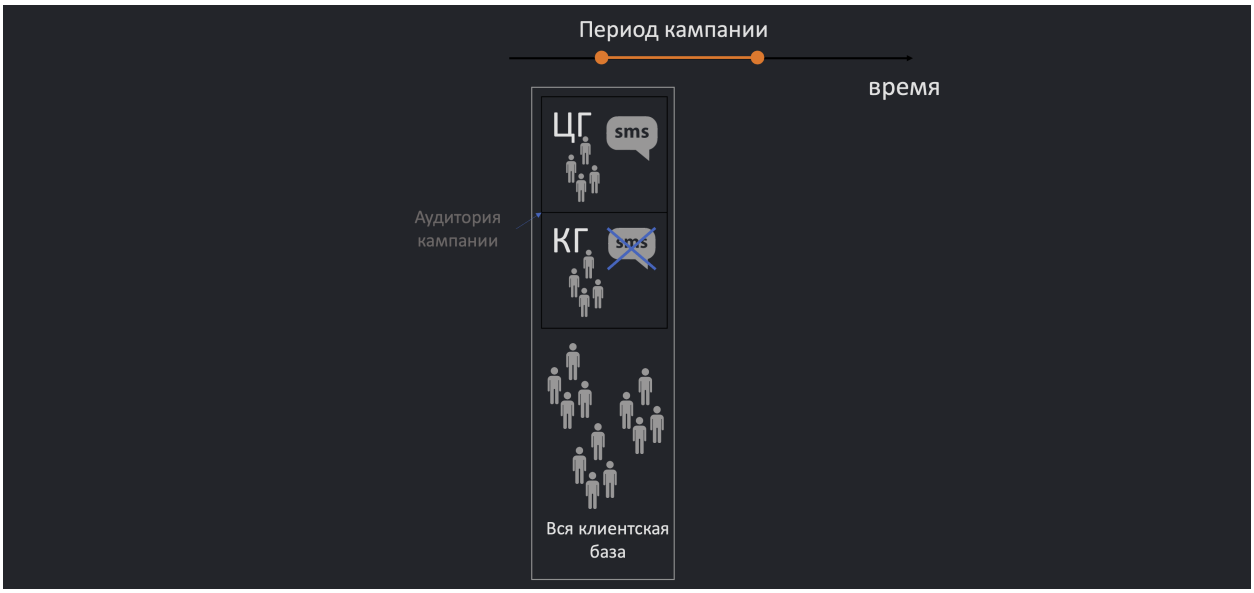
Возможные каналы напоминания:

- SMS;
- Push-уведомления;
- E-mail;
- Слип-чеки;
- Уведомления на сайте или в приложении.

Примеры целевых показателей, на которые мы можем смотреть:

- Число установок, просмотров, покупок и т.д.;
- Выручка;

- EBITDA, т.е. учитываем затраты(на коммуникации, скидки, закупки) — самый сложный вариант.



Указанные показатели рассчитываются за определённый период времени после сделанного предложения.

> Какой может быть бизнес-задача

Типы предложений:

Выбор предложения (или решение не делать предложение) пользователю — Next Best Offer. Примеры:

- Какой величины скидку дать клиенту?
- Какую категорию товаров лучше предложить клиенту?

Уже есть предложение, и нужно выбрать аудитории под предложение. Примеры:

- Кому предложить скидку 10% за покупку от 2000 рублей?
- Кому предложить купить товары для горных лыж?

Возможные ограничения:

- Бюджет на коммуникации;
- Бюджет на скидки;

- Объём аудитории.

> Общая постановка задачи

Обозначения:

X_i — признаки клиента i, описание клиента и контекста.

$T_i \in \mathbb{T}=0,1,\ldots,m$ — вариант предложения, сделанного клиенту (один из вариантов — ничего не делать).

Y_i— целевая переменная на уровне клиента (количество целевых действий, принесённая выручка и т.д. после предложения).

$Y = \sum Y_I$— целевой показатель на глобальном уровне.

Задача:

Не верно: $Y \to \max$ (Y — случайная величина!)

Верно: $E[Y[\vec{X},\vec{T}]] \to \max$, при выполнении ограничений $C(\vec{X},\vec{T})$

Представим, что мы провели маркетинговую кампанию и хотим проанализировать результат:

Конверсия в кампании составляет 20%, в *контрольной группе* (`кг`) — 16%:

👶Неопытный аналитик: отличный результат в 20%! Всё благодаря маркетингу!

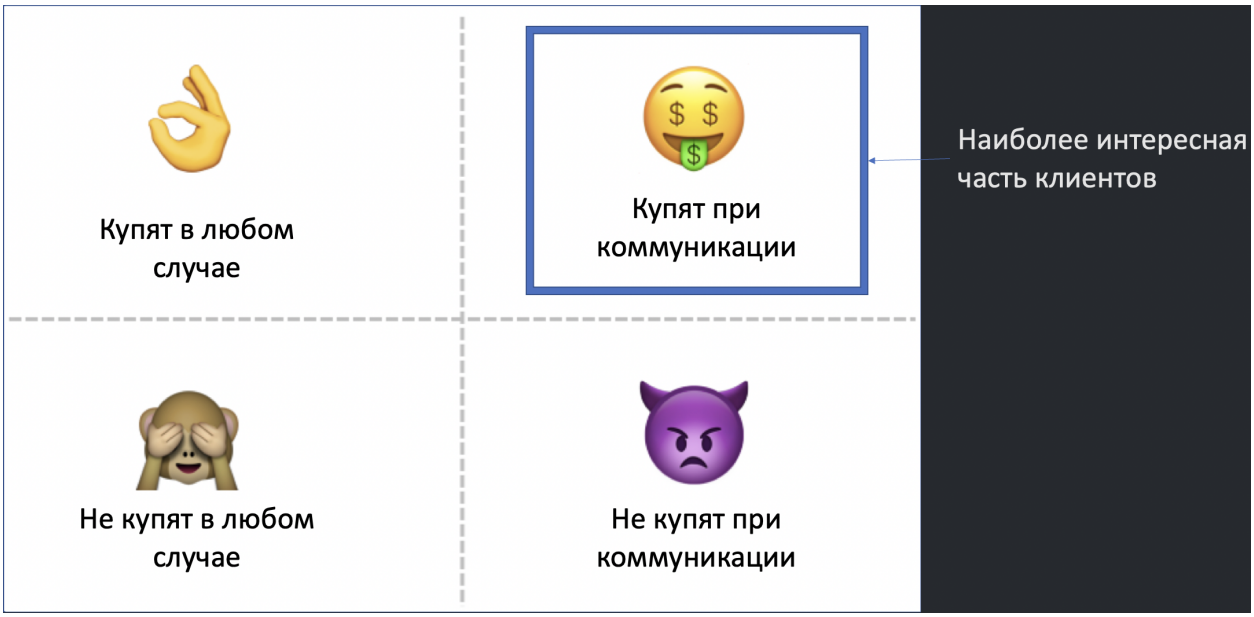
😉Более опытный аналитик: конверсия в целевой группе составляет 20%, а в контрольной — 16%. Значит, мы заработали на конверсии 4% клиентов в `цг` (*целевой группе*).

😎Опытный аналитик: конверсия в целевой группе составляет 20%, а в контрольной — 16%. Значит, мы заработали на конверсии 4% клиентов в `цг`. Но мы потеряли деньги (СМС, скидка и т.д.) на 16% клиентов в `цг`, которые и так бы откликнулись.

> Causal effect

> Типы клиентов

Рассмотрим ситуацию с одним типом предложения: нужно понять, кому сделать предложение, а кому — нет.



Для маркетинга наиболее интересная группа — те, кто произведёт действие только при коммуникации.

> Causal effect

X — признаки, описание клиента и контекста.

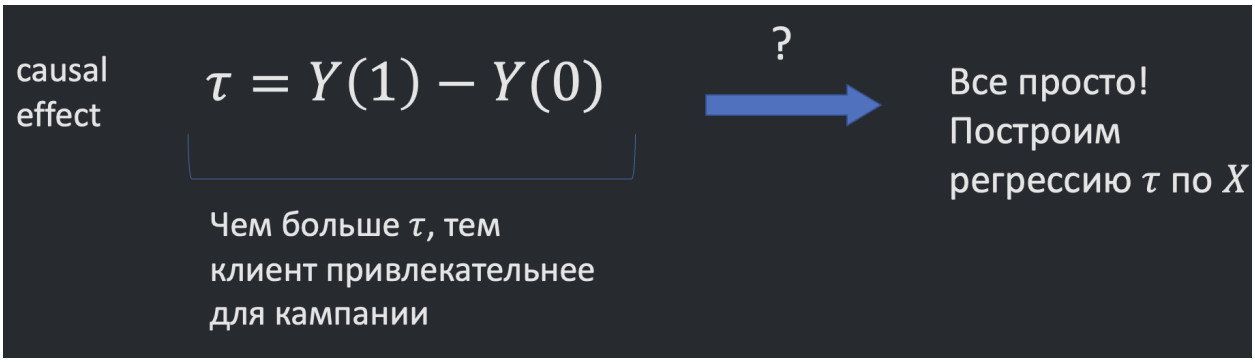
$T \in 0, 1$ — флаг, указывающий, что клиенту сделали оффер.

$Y(0)$ — целевая переменная (реакция клиента) во вселенной, где клиенту не делали оффер.

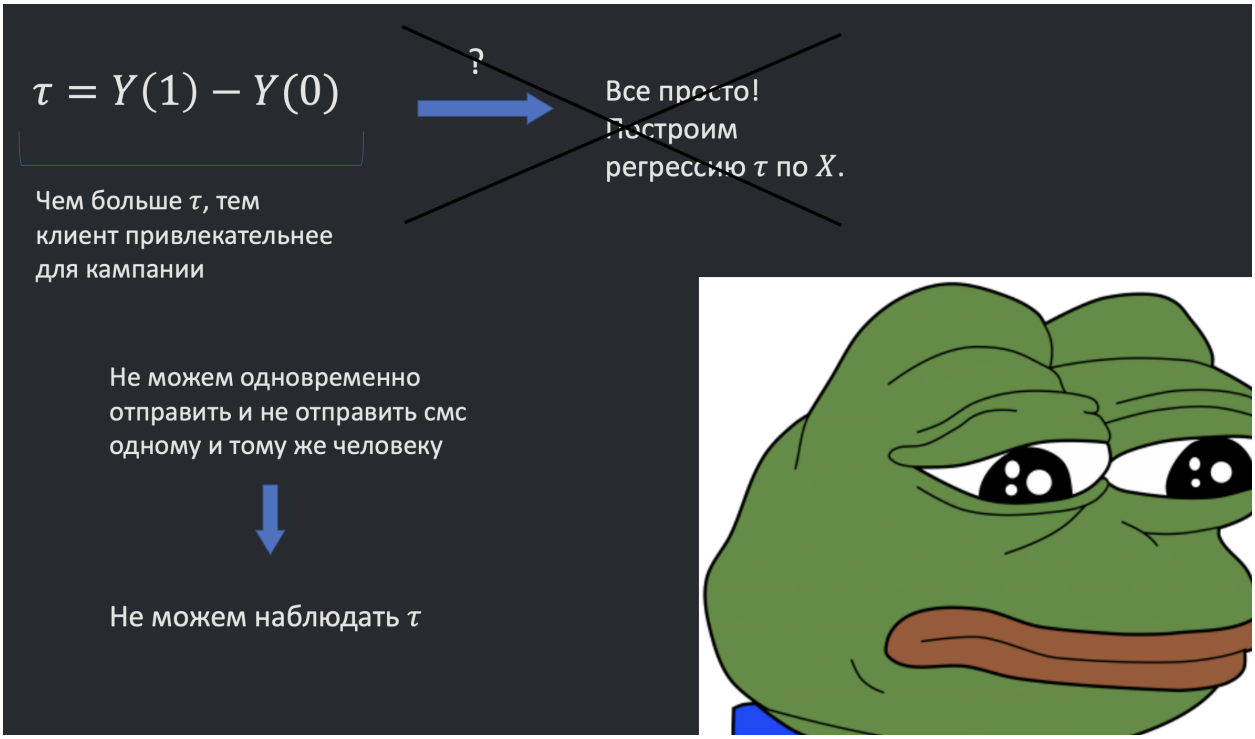
$Y(1)$ — во вселенной, где сделали оффер.

Y – целевая переменная в нашей вселенной, т.е. $Y = TY(1) + (1-T)Y(0)$

τ — `causal effect`, на сколько больше целевых действий совершает клиент, когда ему сделали предложение, по сравнению с тем, когда не сделали.



Кажется, что всё просто: если знаем causal effect, нужно построить регрессию для causal effect от X



На самом деле, это нельзя сделать, так как одному пользователю мы не можем одновременно и выслать предложение, и не выслать :(

> Uplift

Вместо `causal effect` будем прогнозировать `uplift` переменной Y:

$$u_Y(x_i) = E[Y \mid X = x_i, T = 1] - E[Y \mid X = x_i, T = 0]$$

То, насколько больше клиент x_i реагирует с воздействием в сравнении с реакцией без воздействия. Грубо говоря, средний прирост Y при коммуникации.

Наша цель — построить модель `upmodel`, которая будет пытаться прогнозировать `uplift`:

$$u(\hat{x}_i) = upmodel.predict(x_i)$$

Когда модель построена, её можно применить ко всем клиентам из базы и выбрать тех клиентов, для которых предсказание наибольшее.

> Общая схема построения uplift-модели

> Необходимые данные

Требуется история по прошедшим акциям и коммуникациям.

> Как должны выглядеть данные

По каждому событию, когда пользователю делали предложение:

- То, каким был клиент на момент, когда предложение делали (*x*, можно включить контекст и другие данные).
- Флаг, указывающий, в какой группе был клиент (контрольной или целевой).

- То, как клиент отреагировал — сколько было целевых действий после получения предложения (целевая переменная).

ID клиента	Сегмент лояльности	Число покупок за пред. 30 дней		Флаг ЦГ	Число покупок в период кампании
Дима	loyal	9	...	1	3
Анна	uncommited	3	...	0	1
Игорь	potential	6	...	1	0
...			...		
	X			T	Y

Три блока необходимых данных

> Применение uplift-модели

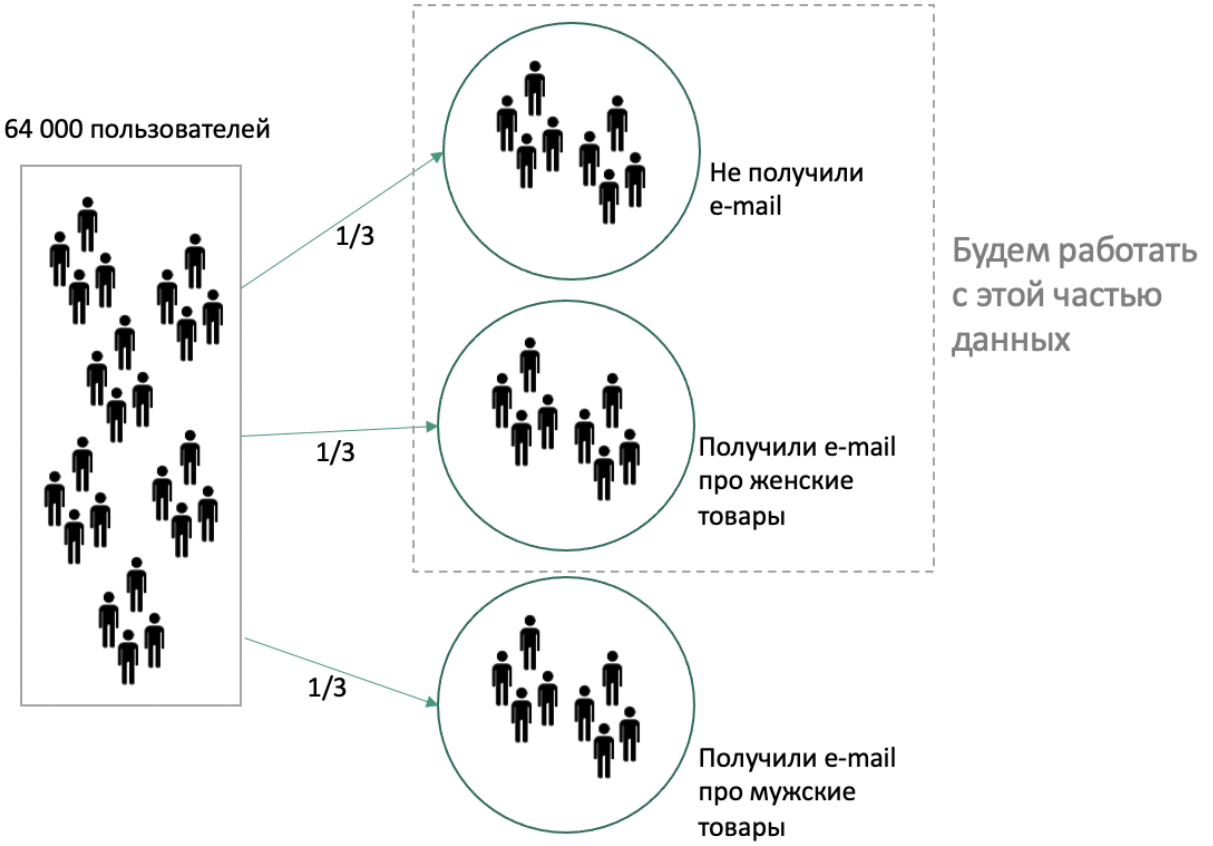
Ноутбук с практикой: скачать.

Импортируемый модуль: скачать.

Берём данные о текущем состоянии клиентов и применяем модель. Флага ЦГ или КГ нет.

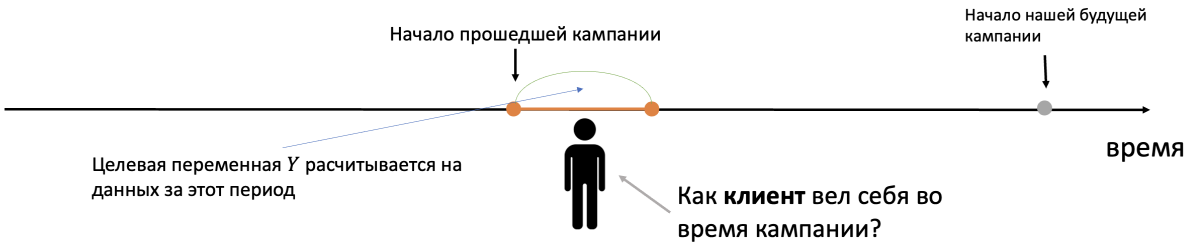
Посмотрим на пример старого соревнования от MineThatData:

В датасете содержится информация о 64000 клиентов интернет-магазина, которые совершили покупку в последний раз в течение 12 месяцев перед e-mail рассылкой.



> Целевая переменная

Должна отражать, сколько пользы бизнесу принёс клиент после получения предложения, т.е. должна описывать поведение клиента после рассылки.



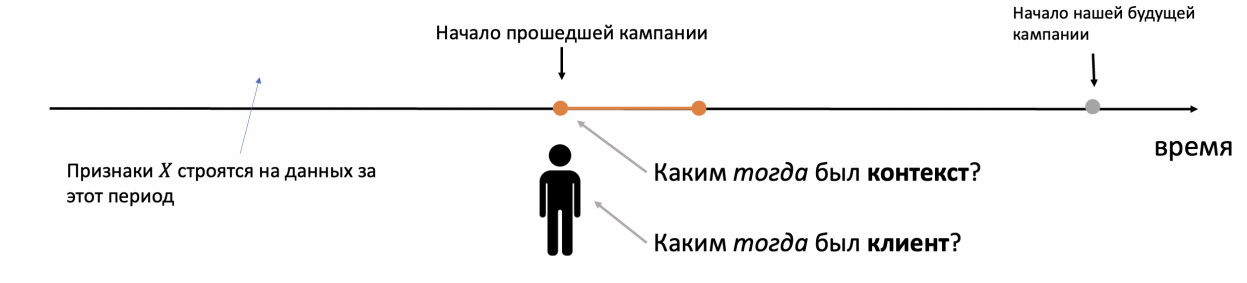
Хотим максимизировать целевой показатель кампании:

$$Y_{total} = \sum_{i \in clients} Y_i$$

Примеры для разных целей по уровню всей кампании и одного клиента:

Уровень кампании	Y_{total}	Общее число чеков	Выручка (РТО - розничный товарооборот)	EBITDA
Уровень клиента	Y_i	Число чеков клиента	РТО клиента	РТО клиента - себестоимость товаров клиента - затраты на смс клиенту - затраты на баллы клиенту

> Признаки



Описание клиента на момент времени до воздействия на него

Примеры признаков по клиенту:

- История чеков — например, число покупок, выручка и т.д. за последние k дней.
- Социально-демографические данные — например, пол, возраст и т.д.
- Географические признаки — например, уровень населённого пункта клиента (мегаполис, город, деревня).

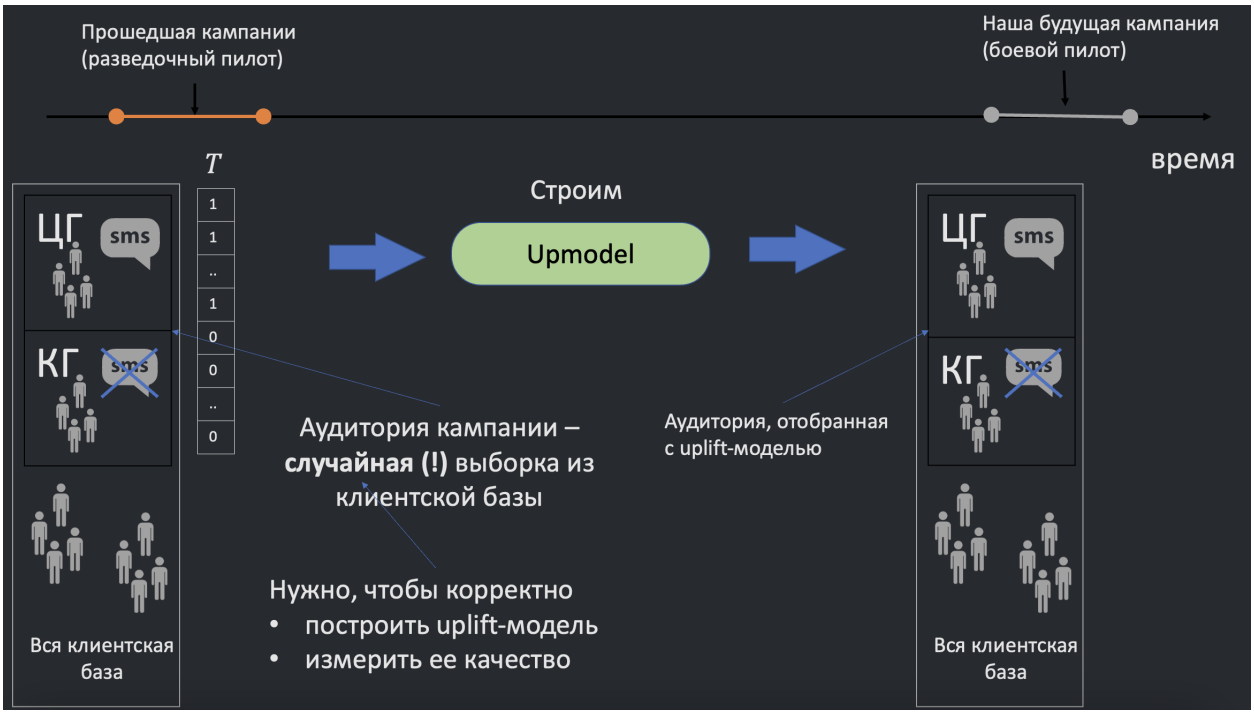
Примеры признаков по контексту:

- День недели;
- Месяц;
- Флаг попадания праздника в период кампании.

> История кампании

До построения модели уже должна быть минимум одна кампания. К ней предъявляются следующие требования:

1. Аудитория — случайная подвыборка из базы (желательно);
2. Одна часть аудитории относится к ЦГ (рассылали коммуникацию), другая — к КГ (не рассылали коммуникацию).
3. ЦГ и КГ одинаково распределены.



Все данные кампании (можем добавить и более ранние) скармливаем алгоритму построения uplift-модели. После этого проводим новую кампанию:

- Выбираем клиентов с максимальным uplift прогнозом;
- Выделяем случайным образом КГ (чтобы оценить коммуникацию в будущем).

Из-за выбора аудитории, основанного на uplift-модели, пользователи уже не случайны. Можно добавить дополнительных случайных пользователей из базы. Эту группу также нужно разбить на ЦГ и КГ.

Добавление случайных пользователей нужно для того, чтобы произвести exploration для получения данных и использовать их в построении будущих uplift-моделей.

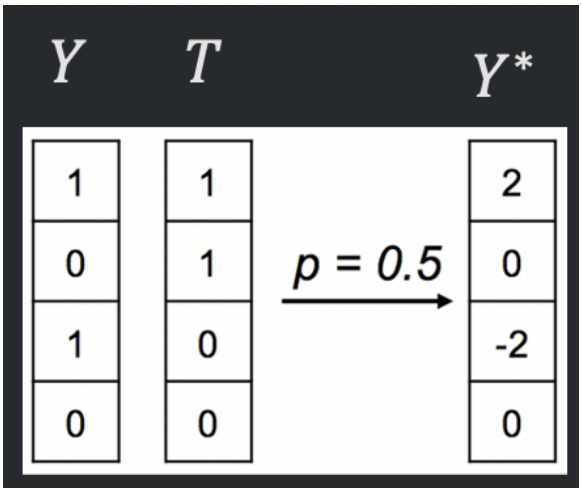
> Некоторые подходы к моделированию

- Обучить одну модель, где флаг T используется как признак. Для клиентов из тестовой группы делаем предикт дважды: сначала подставляя в модель признак `treatment_flg` единицу, затем ноль. `Uplift` будет разницей этих прогнозов.
- Обучить две модели на подвыборках: $T = 1$ и $T = 0$. Потом использовать разность их прогнозов(аналогично первому пункту).
- Трансформация таргета: по имеющемуся Y и T рассчитывается новый показатель Y^* по клиенту, который потом будем использовать как целевую переменную.

$$Y^* = Y\frac{T}{P(T=1 \mid X)} - Y\frac{1-T}{P(T=0 \mid X)}$$

Что при случайном разбиении аудитории на ЦГ и КГ эквивалентно:

$$Y^* = (2T-1)2Y$$



Заметим главное свойство Y^*:

$$E[Y^* \mid X] = P(T=1 \mid X) \cdot E[Y^* \mid X, T=1] + P(T=0 \mid X) \cdot E[Y^* \mid X, T=0] = u_Y(X)$$

Можно надеяться, что прогнозы данного показателя будут сходиться к нужному uplift.

Строим регрессию с таргетом Y^*. Используем loss-функцию, которая даёт схождение прогнозов к мат. ожиданию, например `MSE`.

Прямое моделирование uplift:

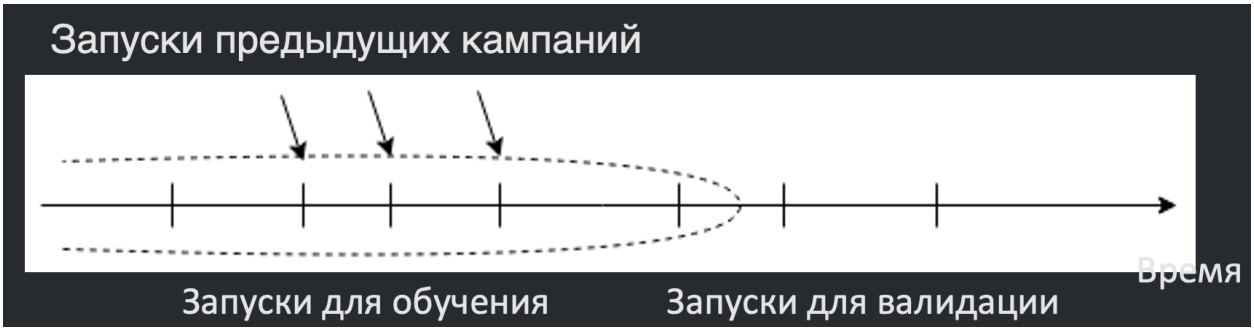
Строим решающее дерево, где перебором всех признаков и порогов определяется порог с максимальным критерием информативности, чтобы образовались как можно более различные по распределению uplift группы.

> Оценка качества

> Схема валидации

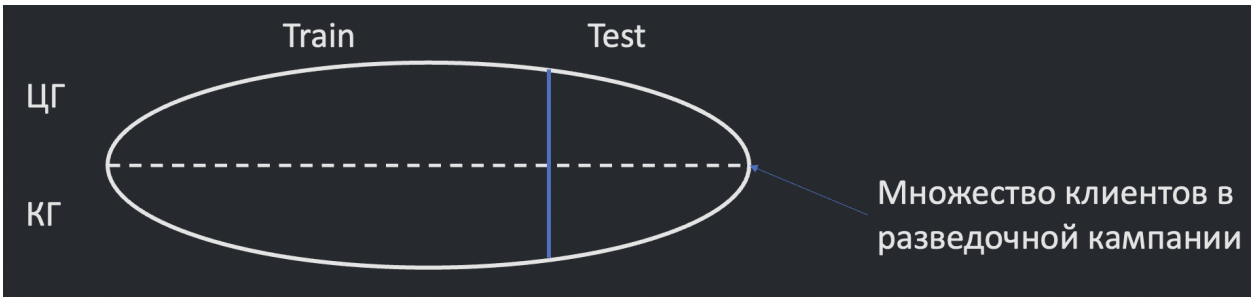
В зависимости от доступных данных можно выделить два случая:

- Есть много старых кампаний.



Часть данных о более старых кампаниях возьмём для обучения, а более новые — для валидации.

- Есть одна старая кампания — не получится симулировать реальную ситуацию.



Все данные требуется разбить на train и test независимо от отнесения клиента к КГ или ЦГ.

> Метрики

Изучим график `Cumulative gain curve` (*uplift кривая*).

Его смысл:

- Применяем uplift-модель ко всему множеству клиентов;
- Затем сортируем клиентов по убыванию прогноза;
- Теперь в каждой точке, соответствующей доле топовых клиентов p, считаем:

$$CG(p) = (\frac{Y_p^T}{N_p^T} - \frac{Y_p^C}{N_p^C}) \cdot (N_p^T + N_p^C), \text{где}$$

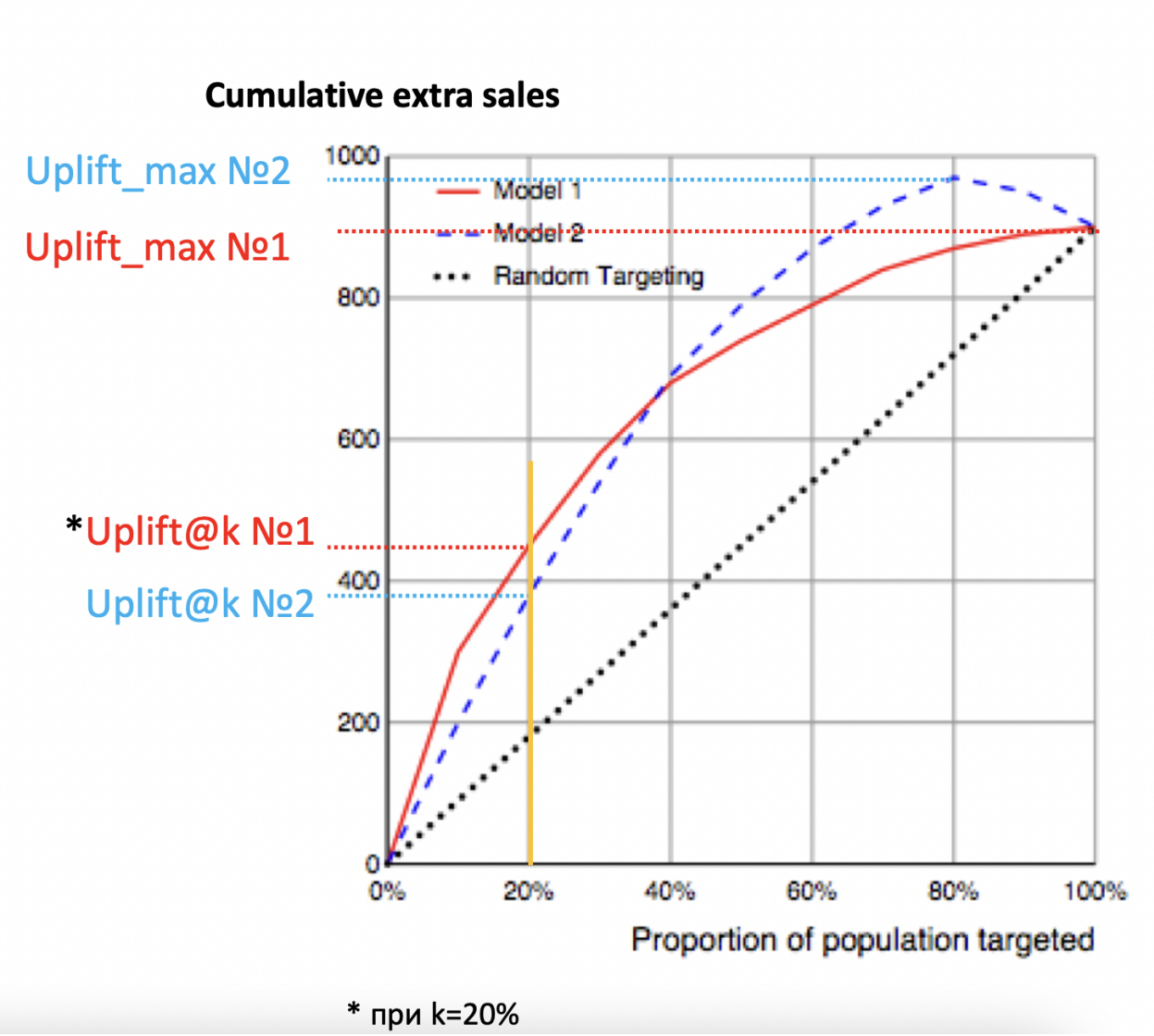
Y_p^T, Y_p^C — сумма по Y для $T = treatment$, $C = control$ (для первых $p\%$ клиентов).

N_p^T, N_p^C — количество клиентов для $T = treatment$, $C = control$ (для первых $p\%$ клиентов).

Первая скобка отвечает за то, на сколько среднее для ЦГ отличается от среднего для КГ.

Второй множитель — количество клиентов в $p\%$.

Точка на кривой — сколько условных единиц (зависит от целевого показателя) мы заработаем, если кампания будет запущена только среди топовых $p\%$ клиентов.



Чёрная диагональ — случайная сортировка клиентов.

Большой наклон в начале означает, что топовые клиенты, по мнению модели, приносят много у.е. Именно это мы и ожидаем от модели: сначала самые хорошие клиенты по прибыли, потом чуть менее хорошие или клиенты с отрицательными значениями. Хорошо, когда кривая выгнута вверх.

> Численные метрики

- `AUUC` (*Area under uplift curve*) — площадь под кривой.
- `Uplift@K` — значение uplift кривой в точке K. Полезно, когда есть конкретные ограничения, например, на количество клиентов, которых можно использовать.

- `Uplift_max` — максимальная точка подъёма графика. Актуальна, когда целевой показатель — деньги, так как позволяет выбрать максимальный потенциальный эффект (синий график — оптимальные 80%).